Description of a Microdata Access System
For Presentation to the Census Advisory Committee of Professional Associations October 27, 2006[1]

Philip Steel*,Arnold Reznek**, and Laura Zayatz***

* United States Bureau of the Census, Statistical Research Division, Rm 3209 FB-4, 20233-9100, USA
philip.m.steel@census.gov
** United States Bureau of the Census, Center for Economic Studies, WP2-206, 20233-6300, USA
arnold.phillip.reznek@census.gov
*** United States Bureau of the Census, Statistical Research Division, Rm 3209 FB-4, 20233-9100, USA
laura.zayatz@census.gov

Abstract

This paper describes a software system being developed by the Census Bureau for access to confidential microdata. This system would allow users to specify a model to be run against the microdata and return model coefficients, descriptive statistics and measures of model fit. The context for its development is presented and we continue with a discussion of issues in its design. This paper is an excerpt from a paper presented at the UNECE conference on statistical disclosure control at the end of last year. The remainder of the paper, primarily a discussion of technical problems, can be found in the appendix. In the period since the original paper was written there has been a hiatus in development while various funding and IT issues have been addressed. Some progress has also been made in solving the very critical differencing problem. Work on the system resumed in September, with the short term goals of accessing ACS data (in addition to the original CPS dataset), adding the ability to specify interaction terms and making preparations to deploy the system in the Ferret environment. Our questions for the committee:

What data sets should we try next? (demographic, economic)

What other types of analysis should we try to include?

Who can help with testing? (data users, confidentiality filters)

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

Microdata publication has been the avenue of choice for data producers to serve the needs of sophisticated data users. Microdata allows virtually any analysis, but it is the only avenue of publication that allows modeling. To produce microdata in the context of a survey that promises the confidentiality of responses, the identity of respondents must be hidden. This has been accomplished by a variety of means, but the staple method is to suppress low-level geography and coarsen other variables until there is ambiguity as to whom in the population a record corresponds. Evaluating whether microdata is safe is problematic and often reduces to whether or not a high quality, identified, external file with sufficient overlap exists or can be constructed. As public data become more and more accessible, and data of all types on individuals accumulate, the ambiguity we rely on to protect the respondent is reduced. At the Census Bureau, we have undertaken a continual review of external data and have reduced detail on our microdata publications as potential problems are discovered. It is not difficult to project that at some point, microdata will be unsafe to publish, or what is publishable will be of low utility. On the one hand, the demand for microdata, both for general research and programmatic needs, continues to grow, and on the other, its separation from identified public or commercial data is harder to maintain. At the conference marking the publication of “Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies” this problem was dubbed “the train wreck” [Doyle et al 2001].

Model servers represent one way out of the train wreck problem. The result of the model is the object of interest, not the underlying data. There are indications that most model results are safe, at least when considered in isolation. The Census Bureau operates a number of Research Data Centers (RDCs) where researchers with Special Sworn Status have access to specified microdata. Model-oriented research is encouraged, and the output that is to be removed from the research data center is reviewed by the on-site employee and sometimes by the Census Bureau's Disclosure Review Board (DRB). The model output that we have examined over the past ten years of operation has been virtually without disclosure problems. Can this be reproduced in an automated system?

1.1 Existing model query systems

Several statistical query systems for analysis of sensitive data allow modeling: the Luxembourg Income Study (LIS), the (US) National Center for Health Statistics' ANDRE, and (US) National Center for Education Statistics’ DAS are the most proven. All three systems require registration and a statement of purpose. ANDRE has some explicit monitoring capability. The LIS web documentation does not mention monitoring of users for compliance, but does in fact do so. ANDRE operates in conjunction with their Research Data Center (RDC) and provides remote access for the RDC's registered users. LIS is often utilized solely by remote access. Several similar systems are under development in the European Union. DAS takes a different approach by offering correlation matrices instead of running a formal routine. For a description of current systems see [Rowland 2003].

2 Preliminary design issues

The Census Bureau funded the development of a prototype model server suited to its needs. It is a proof of concept, using the Current Population Survey (CPS) public use microdata as its test bed, but designed to accommodate other microdata sets, and run SAS; some adaptation of it should be compatible with American FactFinder, our table server system. It would have exploratory capability, be user friendly, run a variety of models and populations, and provide measures of model fit.[2]

The mail systems employed by LIS and ANDRE have certain vulnerabilities. The submitted code could contain blocks that haven't been anticipated. This could include undocumented options and procedures, code that interacts with the operating systems, new procedures, or convolutions to get around procedures that are explicitly banned. Such a system can work in a monitored environment, where the users and their programming style are under observation, but staffing and user registration become an issue.

A web-based system, where code is built to user specification, avoids many of the problems associated with mail-based servers. This approach is termed an "enabling" system in the programming community: activity is restricted to what has been designed into the system, rather than proscribing certain activities or procedures and allowing all others, i.e. "disabling". It solves the problem of understanding the code being submitted by users and evaluating the output for disclosure. The enabling approach also has its pitfalls: it requires a user-friendly interface and is limited to only those statistical models that it has been programmed to construct. We have opted for an enabling system, since the monitoring problem of the disabling approach seems inescapable and the confidentiality problems in an enabling system can be addressed as capabilities are added to the system.

A computerized system generates its own set of problems, some quite different than what is encountered in the RDC environment. In particular, there is a problem of accumulated results and the inferences one can make from them. This is a variation of what is commonly referred to as "the subtraction problem". In current practice, preliminary results never leave the center and we rely on the RDC administrator or the DRB to recognize when final model results are too similar or when the results are not germane to a researcher's purpose. The process of producing a release from an RDC may involve months or years of work; a computerized system can produce the same volume of results in minutes. With an enabling system the problem of accumulated results is bounded (since the range of queries is known and limited) and can be confronted more directly.

2.1 The Current Population Survey

The system is meant to have general applicability, but we were also interested in seeing the extent of the difficulties that would be encountered on large complex data and what aspects of the system required survey specific programming. The public use dataset for

[2] Synectics is the developer of the system.

CPS was selected for a test bed, specifically the March 2000 supplement. The CPS is the United States' longest running survey, tracing its roots back to an effort to measure unemployment during the Great Depression. Microdata from 1994 on are freely available at http://www.bls.census.gov/cps/. The March supplement focuses on demographic data and widens the applicability of the test. The data are well documented and have carved out a place both in current research and as an educational tool [Berndt 1990]. Using a public-use file allows us to involve more people in the testing of the software system, without risking confidential data. The data are topcoded and have a prepared geography, and any issues with categorical variables have already been resolved. The system's dependency on this kind of preparation must, at some point, be evaluated.

2.2 Non-confidentiality problems in design

As with any data tool, the model analysis system has to present the options of a complex task in a simple to use format. Hierarchical data structures, such as geography and the relationship of the individual in groupings such as household and family, affect the ultimate structure of the queries and the confidentiality problems encountered. Yet they are also the items frequently of interest to the user. The CPS is rich in detailed financial data with a data dictionary that is 127 pages long. Variable descriptions must be incorporated into the instrument and often determine the role of the variable. A variable may play several different roles or, for confidentiality reasons, be barred from playing a particular role; for instance a poverty indicator (which combines income thresholds with household size) can be available in the exploratory phase, cannot be combined with income in universe formation, but can be available again in the analysis phase. The handling and display of the survey metadata is a large piece of the overhead for this project.

2.3 Confidentiality strategy in design

The design can be divided into five moderately distinct sections: data preparation, data exploration, universe definition, model statement and results. Because the basic strategy is to "enable", most restrictions will be passive from the user's point of view. They simply have no facility to engage in risky behavior. We try to avoid active restrictions, where the user makes a choice, it is evaluated, and possibly denied. Active restrictions lead to frustration, particularly if the evaluation comes later in the process or otherwise generates delay.

Data exploration will initially be rudimentary—allowing the user to make only a general examination of the data, sufficient to inform decisions for constructing a model. More descriptive or expressive data exploration can be permitted once the confidentiality requirements are known to be effective.

Universe definition, the restriction to the user's desired population, is more directly involved with determining the parameters of the modelling system and may be the most difficult to accommodate. This will be the substantial focus of the initial development effort. The universe definition stage is equivalent to a "coarse" table server.

While restrictions must be imposed on the model statement, those restrictions address some known very specific problems. These constitute a very small fraction of possible models. The user may never encounter those restrictions associated with the model statement, with the exception of an initial ban on large, fully saturated models [Reznek 2003].

The estimates for the model are derived directly from the data in as much detail as the collection and preparation can afford. The values on which they are based may differ from what is encountered in exploration and what is available in the universe formation stage, where the user may encounter a synthetic analogue or a recode. Diagnostic statistics require considerable care and can pose a substantial disclosure risk. For instance, residual values are record level data and the values they are based on are easily recoverable. Diagnostics may also address outliers. Where diagnostics are risky, a synthetic approach will be employed [Reiter 2003]. The emphasis will be on the ability to obtain model results without noise or bias from confidentiality restrictions. Diagnostics will be approached more conservatively and can be the subject of future improvement.

3 Future Application

The degree to which the architecture of the program will support a shift of datasets or an expansion of models is yet to be determined. We would like to add a user selection of the absorbing dummy category in the simple regression procedure. The presentation of long variable lists can be improved. Usability of the tool, independent of confidentiality concerns, is crucial to supporting its further development.

The model server tool is ideal for instruction. Because no programming knowledge is assumed for the user, students with limited programming or even statistical expertise can learn basic regression analysis on live data. For surveys where microdata are a subsample (Decennial Census, ACS), the model server could be pointed at the nonsampled portion and provide model results for data with very little topcoding. For surveys that have developed replicate weights for variance estimation that are unavailable to the public, the model server system may allow results to be evaluated with those weights without compromising the detail (usually geographic) which makes the weights themselves risky.

4 Conclusion

We have specified a query system with parameters that can be adjusted to tighten or relax confidentiality requirements. We welcome any suggestions or justifications for particular settings of those parameters. We hope we have taken a step forward by producing a prototype system that affords some of the advantages of microdata in an environment that is safer than exists with standard microdata. The system described is designed to prohibit the user from reconstructing the underlying microdata. By this we mean primarily that the record-level organization cannot be recovered. The long-term problem with microdata is the number of variables that overlap with external data and, if successful, this tool gives an alternative that denies application of record linkage machinery.

Appendix

5 Data exploration

It is important that users be able to examine univariate and bivariate distributions before being asked to specify a model. The user may also wish to confirm some aspect of their result with a simple table. The exploratory capability should include such tables for most or all categorical variables, at the user specified geographic level. The exploratory tabulation applies to the entire population, not necessarily the population the user is studying. The confidentiality requirement for this facility is on the data preparation. The preparation must support exploratory two-dimensional tables; for CPS we have preparation primarily through the CPS geographic designation, where no designation shows fewer than 100,000 population. For numeric variables we can offer a categorical analogue or a synthetic representation. The categorical analogue may be a simple indicator or something more detailed. A synthetic representation allows us to extend the exploratory capability to a display of plots. This is particularly suited to displaying transformations of numeric variables. Transformation will be limited initially to log, square and square root, but this list may be expanded at a later date [Reiter 2003].

6 Universe formation

Universe formation, or subsetting down to the model population, is an overt confidentiality problem. It gives the count of the population defined by some set of conditions, i.e. it is gives a cell in a table of counts. With complete freedom to vary the conditions, any table could be constructed, including cells of size one. But the ability to run a model on one additional observation may allow the reconstruction of the record in its entirety. See [Cox 2004] for the construction of dependent variables. Disclosure avoidance techniques used on tables could conceivably be used, but the more sophisticated ones would be difficult to apply. Disguising the number of observations in the user's desired population gets into sequence, retention and additivity problems. For example, controlled rounding is attractive for this sort of problem, but to work in the context of the server it would have to be performed consistently on all possible tables and the appropriate rounded value would have to be presented to any user whose population corresponds to that particular cell. Later values produced by the model would have to be consistent with the number presented when the universe is initially defined. Cell suppression, or in this context, a rejection of the user's universe selection, can also lead to an open-ended problem.

The problem is not insurmountable, however. Models are usually run on fairly substantial populations, and hence equivalent to moderately large table cells. We will assume (or rather, require) that the model universe will have at least 75 observations. The magnitude is such that we should be able to guarantee (in the data preparation) that the balance of the table cells does not fall below a count of 4. The task then becomes to verify that the user is attempting to model a reasonable sized population and the balance is not a confidentiality problem.

6.1 Numeric variables in universe formation

For categorical variables one would use an equality condition variable to extract the subset on which the model is to be run. For example, head of household=1, Hispanic=1, and educational attainment=44. For numeric variables, like income, the condition would be defined in terms of an inequality. Note that the system also allows for “or” conditions, but the simplest case is sufficient for illustration. For example, all households with total income greater than or equal to 17,000. However, for numeric variables the underlying data have full detail. By incrementing the value, or cutpoint, it would be possible to define a universe for all households with total income greater than or equal to 17,001. This could contain just one additional observation. Comparing coefficients of the model run on both universes may indicate the characteristics of that isolated record. By going through a progression of models, it may be possible to reconstruct the entire microdata record. How can one prevent differencing of this type, but still provide some facility for using numeric variables in defining a universe?

Clearly the set of cutpoints must be pre-determined. We would like the points to be evenly spaced, with enough distance between points so that there is a reasonably good chance that the user will not run into a denial based on the "at least 4 observations" rule. When numeric data are being rounded or presented in tables, the schemes used are often ad hoc; but they share a property of graduation, so that the average difference between the true value and the rounded value is proportional to the magnitude. Rounding by 10 up to 200, by 100 up to a 2000 etc. Examination of the CPS data on income reveals a great deal of rounding by respondents, following a pattern that is also responsive to magnitude. Generally the data spikes initially at values divisible by 5, then later by 10 then 50, 100, 500, 1000, 5000, 10000. The peak of observations at 30000 represent some values being rounded up, and some being rounded down. The rounding scale is also varying, with some people rounding 32,500 to 32,000 and some rounding all the way to 30,000. The initial set of cutpoints for the long list will consist of a compromise between an attempt to evenly distribute the data and the incrementing between peaks that occur naturally in the data. The cutpoints will be calculated using the natural increments with the shift in increments sensitive to a threshold on the number of observations. Cut points will be offered at 50, 100, 150 until the number of observations between falls below our parameter, at which point the increment will increase to 100 to continue until it fails to capture enough observations, bump the increment and so on. The set of cutpoints distributes fairly evenly across the data, can be adjusted in the testing phase, and follows a scheme that is similar to the clustering already occurring in the data.

On the high end, the cutpoints should not exceed the value one would use for a topcode in a microdata publication; that is, a half percent of all observations should be above the topcode or 3 percent of the non-zero observations. Note that this restriction is applied to universe formation. No restriction is currently envisioned on the use of large values in the model input. This feature of the system is not applicable to the current test. The CPS is already topcoded and already published as microdata. Since the record structure is already known, we cannot additionally allow access to large values in universe formation and, perhaps, even in the modeling phase.

Note that the burden is shifted from confidentiality to usability. More granularity in the cutpoints leads to more rejected universe formations. The optimum setting on the threshold for the minimum number of points between cutpoints is a function of what “large” is in the context of users running models on “large” populations. Our initial set of cutpoints was generated to accommodate restrictions where the categorical variables define a population approximately one tenth of the CPS universe. The numeric variable adds a further restriction, but one that is unlikely to violate the four observation rule, by design. This long list of cutpoints is appropriate only for users that require a fairly exact threshold on a numeric variable and have few other restrictions for the model's population. The loss in precision is not as great as it might seem. More hinges on whether a natural rounding point is included or excluded and how much of the contribution to that point comes from respondents rounding up or rounding down, than hinges on the distance from the desired point and the cutpoint actually available. The long list would remain fairly close to a rounded version of the variable, at least for modest increases in the threshold.

For users where categorical restrictions reduce the population to less than one tenth, a less detailed list of cutpoints will be available--a short list. The conditions defining the population can be viewed as a cell in the table of counts obtainable by varying the conditions. The dimensionality of that table is the number of variables involved in the conditions. The size of the table, then is the product of the number of conditions associated with the variables. Our strategy for the short list is that its size will be the square root of the size of the long list and thus two short list variables will be roughly equivalent to the long. The short list is fixed and is a subset of the long list.

6.2 Indicators in universe formation

In addition to short and long lists, an indicator will be available for zero values and perhaps other categorical characterizations. For example, 1 or –1 may have a special meaning. Some variables have a significant range of negative numbers. Whether these are allowable needs to be related to the threshold used in 4.1. An outstanding issue is how to handle derived codes, like poverty. Poverty can be expressed as a function of income, size of household and number of children present. Poverty could be used to subdivide a cutpoint range, since it includes other dimensions implicitly. For such codes an active restriction must be employed, and such restrictions are dataset specific.

7 Confidentiality for the Model Statement: Interactions and Dummies

Disclosure risks may arise from the use of regression models, particularly in the standard linear regression model estimated using Ordinary Least Squares methods as well as in logit and probit models (which use binary (0,1) dependent variables), and other Generalized Linear Models [Reznek 2003, Reznek and Riggs, 2004]. The risks in regression models that contain continuous variables are small if the overall sample is large enough to pass tabular disclosure analysis. Risks are most apparent in models that contain dummy variables as independent variables. Coefficients of models that contain only fully interacted sets of dummy variables on the right-hand sides can be used to obtain entries in cross-tabulations of the dependent variable broken down by the categories defined by the dummy variables. That is, from the disclosure avoidance point of view, these models are equivalent to tables. We will conservatively bar interactions involving 4 or more variables and fully interacted models of 3 variables. It also seems sensible to keep users from specifying an over-determined or nearly over-determined system. We will use either a fixed cap of around 20 variables or a parameter dependent on the number of observations.

In addition, each dummy category should have at least 20 observations. For dummies meeting this threshold, its estimated coefficient will be shown. For dummies failing the threshold they will be absorbed into the constant term along with the last term of the dummy. The choice of the absorbing term will not be available to the user, though this is a product of programming rather than confidentiality constraints. Menuing for such a choice requires an additional population of metadata and some, as yet unresolved, division of tasks between the query build and the query execution. Our initial procedure, "Proc Reg" does not support a full-fledged "absorb" statement, though enabling it for procedures that do, should be relatively straightforward.

8 Confidentiality for Model Results

The restrictions in the formation of the model statement are intended to guarantee that the coefficients can be presented without further restriction. Most summary statistics on residuals can be displayed safely. Our efforts on presenting measures of validity have been focused on devising a way to display residuals. Examination of residuals is a frequently used method for evaluating models, particularly in the early stages of development. Of course, the actual residuals allow the recovery of the underlying values and cannot be presented to the user. Synthetic residuals, provided they convey the same information to the user, are an effective substitute. In the test on the CPS public use data, the synthetic residuals are being presented side by side with the actual residuals to determine if they are adequate.

Density estimation will be done by the SAS KDE routine. The output from this allows us to generate a random set of points with approximately the same density as the original residual data. The random number generator must have a fixed start, since repeated application of the KDE procedure should show convergence to the original set of points. KDE is also sensitive to outliers so the synthetic data are topcoded at four standard deviations from the mean. The number of topcoded values will be provided to the user. A variation on this procedure can be used to generate two-dimensional plots. The procedure for generating synthetic residuals can be adjusted in two ways. The endpoints used by the procedure can be altered. Also the number of grid points can be adjusted. Both may have some impact on the effectiveness of the procedure.

9 Data Preparation

The weaknesses of model servers are strongly related to the problems encountered in table servers. The capability to restrict models to particular populations yields counts of the population. Varying the restriction parameters enables the user to construct table margins or cells. For some models, the estimates of coefficients in a model are equivalent to groups of table cells. Tabular disclosure problems of this type are found in virtually any publication form--a model server does not solve them. What we can hope to accomplish is to use the model server to restrict the table server problem to a manageable dimensionality and to prevent the reconstruction of individual records and most numeric values. That is, we will assume that the data at which the model server points are sufficiently prepared to safely allow publication of most lower dimensional tables on the available geography. This is dependent on the population, sample size, some regularity in variable categories and the creation of an appropriate geography.

The strictest standard for microdata is k-anonymity. For all variables thought to overlap with external files, there are at least k members of the population displaying any combination of characteristics present in the microdata. K-anonymity is usually considered with respect to a limited set of variables and frequently must make the poor substitution of the sample population for the full population. The model server setting introduces a substantial variation. Because the primary function of the server is to hide the record structure and that record structure can be recovered by subtraction for susceptible records in the universe formation stage, the anonymity that is desired is with respect to the sample population. If a record is unique with respect to three variables in the CPS it does not matter what multiplicity it has in the population ... its uniqueness on those variables can be used to recover the rest of the record. The relationship of the three variables to external data is not of consequence. That is, the anonymity we desire is not restricted to key variables, but rather the larger set of variables made available in universe selection. However, in this set only combinations of the dimension allowed in the universe formation need be considered. Uniqueness on six variables is not a problem provided the user cannot specify that combination for a universe. This is critical because it is at the higher dimensions that the computational complexity in standard k-anonymity begins to kick in. K-anonymity with respect to a four-variable key (quasi-identifier) is a computationally feasible problem. K-anonymity with respect to an eight variable key is not currently feasible [LeFevre et al 2005].

It is also worthwhile to note that if the tabular disclosure is not direct but rather derived by inference, it may not allow the isolation of a record by subtraction. You may know that there is a unique record with four particular characteristics but are unable to separate it from other records. Whether tabular disclosure poses a risk for model results is an open and difficult question.

References

Berndt, Ernst R., *The Practice of Econometrics: Classic and Contemporary* (1990) Addison-Wesley.

Cox, L.H., "Confidentiality Problems in Statistical Database Query Systems," *Research Directions in Data and Applications Security* (C. Farkas and P. Samarati, eds.) (2004) Kluwer.

Doyle, P. et al., editors, "Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies" (2001) North-Holland.

Harris, K.W., Gambhir, V., "National Center for Health Statistics' Research Data Center", Proceedings of 2004 FCSM Statistical Policy Seminar (forthcoming). See also http://www.cdc.gov/nchs/r&d/rdcfr.htm

LeFevre, K., DeWitt, D., Ramakrishnan, R., "Incognito: Efficient Full-Domain K-Anonymity", procedings SIGMOD 2005.

Luxembourg Income Study web site: LISSY V Job Submission Instructions, section A.2 see also http://www.lisproject.org/introduction/userform.htm researcher request form

NCES' Data Access System web site: http://nces.ed.gov/das/

Reiter, J. (2003), "Model Diagnostics for Remote Access Regression Servers" *Statistics and Computing*, 13, pp. 371-380.

Reznek, A. (2003), "Disclosure Risks in Cross-section Regression Models" Proceedings of the Section on Government Statistics, JSM.

Reznek, A. and Riggs, T. (2004), "Disclosure Risks in Regression Models: Some Further Results". Proceedings of the Section on Government Statistics, JSM.

Rowland, Sandra (2003), "An Examination of Monitored, Remote Microdata Access Systems" from the National Academy of Science's Workshop on Access to Research Data: Assessing Risks and Opportunities.